[Reprinted from FOOD TECHNOLOGY, 1958, Vol. XII, No. 5, Pages 231-236]

Sensory Difference Tests[a]

(Manuscript received September 14, 1957)

David R. Peryam [redacted]
Quartermaster Food and Container Institute for the Armed Forces, Chicago, Illinois

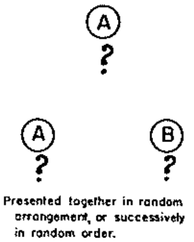
THE PURPOSE of all testing is to establish differences. Since this paper does not propose to cover all sensory testing, and since boundaries are needed, it will be concerned mainly with certain methods which are commonly accepted as difference tests. First, however, let us consider some general characteristics of difference tests and attempt to establish a frame of reference for understanding the particular methods.

Characteristics of difference tests. Difference testing usually means that we are interested in "difference as such," i.e., any kind of difference, without regard to its nature or direction. For example, in the flavor quality control of a product, if strict identity with the product standard is desired, one uses a difference test; but if all that is wanted is assurance that each production lot is at least as good as the standard, some other approach, such as preference, is used. Results of difference tests are easily interpreted—either the panel discriminated or it failed to discriminate. A related characteristic is that they are usually based on statistical models having a known theoretical distribution of responses that would result from chance if the panel were completely unable to discriminate. A third characteristic is that the subject's task is completely defined in the experimental situation, so that all he has to bring to the test is his sensory capacity, plus some native intelligence. He does not have to remember standards or bring his past experiences to bear, but needs only to compare and contrast the stimuli presented. But this "general principle" is often violated—and to good advantage. A fourth characteristic is that the range of the responses allowed is usually quite limited. A test unit is set up so that the subject makes a single choice which is either "correct" or "incorrect." It is "correct" if it supports the hypothesis that a difference exists, and "incorrect" if it does not. In reviewing the test methods these "essential" characteristics will serve as reference points.

Statistics. It would mean little for just one or two people to choose correctly upon one or two occasions. It could easily happen by chance. But how many people, on how many occasions, must be correct before we may infer that we have a real difference? Here statistical interpretation is needed as a means of correcting for chance. The statistical model is a simple one and several methods of analysis are possible. The *t* or binomial distributions are perhaps the most often used; however, all methods give similar results. Charts or tables showing the significance of results can be easily constructed *(8)*. Since the statistical analysis has been adequately covered elsewhere *(12)*, it will not be treated in detail here; however, two important factors determining the significance of a test result should be noted. As the difference between the obtained percentage of correct responses and the chance percentage increases, and as the number of responses increases, the probability that the result could have happened by chance decreases.

TEST METHODS

Figure 1 shows the triangle test. This is the best-known form, most used and most written about, and perhaps also the most useful. The method was first published in 1946 by two Danish statisticians, Helm and Trolle, who described it in connection with flavor quality control work for a brewery *(9)*. It was developed independently by an American distiller in 1941 for use in research and control *(16)* and probably was used still earlier by others.


A
?
A
?
B
?
Presented together in random arrangement, or successively in random order.

JUDGMENT: Selects the "different" sample

STANDARD: None designated, but criterion is within test

NUMBER OF UNKNOWNS: 3

CHANCE PROBABILITY: 1/3

Figure 1. Triangle test.

Here a test, or test-unit, consists of 3 samples. Two materials, *A and B*, are tested to determine if they differ. The subject is given 3 samples, two *A's* and a *B*. All three samples are coded, so that as far as the subject is concerned they are "unknowns." However, he knows that there are two of one and one of another, and he is told to pick the "different" sample. Thus, there is no designated standard, but the criterion for judgment is included in the test, i.e., it can be developed on the basis of these 3 samples alone. Either of the two test materials may be duplicated, and sometimes the practice is followed of alternating the material to be duplicated from one subject to the next. Usually the samples are presented simultaneously in random arrangement, although they can be presented successively in random order. Finally, let us note that since there are three unknowns, chance probability is one-third, i.e., on the average one response out of three would be correct even if the materials were identical.

The paired test (Figure 2) lacks one of the general characteristics of difference tests listed above. The standard is not immediately present, but is a subjective, predesignated one which the subject brings to the test situation. The subject is given 2 samples, again either successively or simultaneously, and is told to select one of the two according to the given criterion. To make this test effective the criterion must be such that it is understood and reacted to by all of the subjects in the same way, e.g., saltiness, sweetness, spoiled flavor, hardness, or some other familiar property. If not, the test becomes merely a mat-

[a] This paper reports research undertaken at the Quartermaster Food and Container Institute for the Armed Forces, and has been assigned number 780 in the series of papers approved for publication. The views or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the views or indorsement of the Department of Defense.

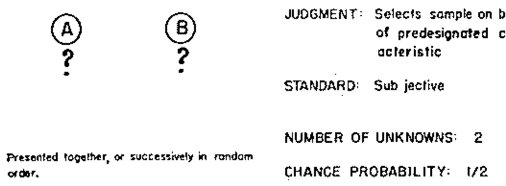

A
?
B
?
JUDGMENT: Selects sample on basis of predesignated characteristic
STANDARD: Sub jective
NUMBER OF UNKNOWNS: 2
Presented together, or successively in random order.
CHANCE PROBABILITY: 1/2

Figure 2. Paired test.

ter of voting, and one cannot interpret a negative result to mean that the panel failed to detect a difference but only that they failed to agree. Since there are 2 unknowns, the chance probability of a correct answer is one-half. This test is simply a special application of the psychophysical method of paired comparisons which has been standard for more than 75 years *(7)*.

Figure 3 represents another paired test, but with the standard present. This is the duo-trio test, originally developed by the Joseph E. Seagram Quality Laboratory in 1941 *(16)*. One of the paired samples is identified and presented first. Then the subject gets another *A* and a *B* as unknowns, in random order. The time interval between samples can be varied as desired. The subject's task is to pick the ''different'' sample. Again, as in the triangle, the standard is back in the test situation, and we even show the subject what to look for. Since we have reduced the unknowns to two, the chance probability of being correct is one-half, rather than one-third as in the triangle.

Figure 4 is another variant of the paired test, which has been given the name, ''dual standard.'' Here both *A* and *B* are presented first as knowns, then they are given as unknowns and the subject has to identify them. This test seems easier than the others because the subject is given a complete ''dry run'' and one might expect a greater degree of success. However, there are disadvantages, particularly in taste tests where tasting the additional samples seems to be more confusing than helpful. For odor testing it does appear to give better discrimination. This procedure was also developed by the Seagram Quality Laboratory *(16)*.

A new kind of situation is represented in Figure 5. If we can use two *A*'s, what is to prevent us from using three, or even more? Nothing, except that there would be no advantage when there are only two materials, and it would increase the complexity of the test. But if the standard is not completely uniform, this multiple standards test can be used. Assume that we want to test the unknown, *B*, for difference against the nonhomogeneous standard, *A*. Of course, we could test *B* individually against A_1, A_2, etc., but since we already know that the *A*'s themselves differ significantly, the results would not be very meaningful. In the multiple standards situation *B* is presented along with a series of *A*'s, with the instruction not just to pick the ''different sample'' but to try to find the one which is ''most different,'' i.e., does not belong to the family. No standard is designated as such, but the criterion for judgment can be established from the test samples. Since all samples are unknown, the chance probability of success is one divided by the number of samples in the test. This method was developed at the Quartermaster Food and Container Institute for the Armed Forces, Chicago, for use with odor stimuli. It is not well suited for taste testing because of the number of samples involved.

Figure 6 shows another less-well-known form. Here we are concerned not with a single judgment, but with a whole series. Two names are applicable—''single stimuli'' because of the method of presentation, and ''*A*-not *A*'' because that name was applied at Brown University where the method was studied extensively *(17)*. *A*, the standard, is first presented to the subject several times until he feels he can recognize its flavor. Then the subject is given a series of samples, any one of which may be either *A* or *B* (''not-*A*''), and he must determine which it is. An alternate approach is to present both *A* and *B* first as knowns and then to continue in the same way. The time interval between samples should be somewhat longer than for the other tests—40–60 seconds. A randomization method is used to determine for each presentation whether the sample shall be *A* or *B*. This means that the number of *A*'s and *B*'s presented in a given series may not be equal, although the proportion is a function of a random variable with a ratio of 50-50. Furthermore, the subject must be aware that the order and proportion are determined randomly. Here the standard is of particular interest. The test starts out with the standard or standards physically present and designated, but since there is no checking back, a memory criterion must take over. In long series of samples the criterion may be reinforced by presenting *A*, as a known, periodically.

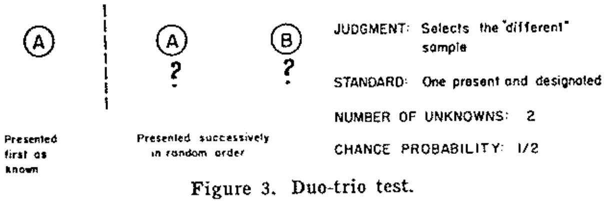

A
A
?
B
?
JUDGMENT: Selects the "different" sample
STANDARD: One present and designated
NUMBER OF UNKNOWNS: 2
Presented first as known
Presented successively in random order
CHANCE PROBABILITY: 1/2

Figure 3. Duo-trio test.

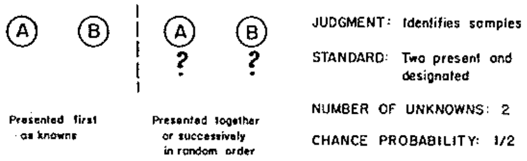

A
B
A
?
B
?
JUDGMENT: Identifies samples
STANDARD: Two present and designated
NUMBER OF UNKNOWNS: 2
Presented first as knowns
Presented together or successively in random order
CHANCE PROBABILITY: 1/2

Figure 4. Dual standard test.

Note that this test is more efficient than any of the others from the standpoint of number of responses for a given amount of tasting. A judgment is obtained for each sample instead of for each set of two or three. The chance probability of a correct response is one-half, but now the subject is selecting one of two possible answers, rather than one of 2 samples.

Figure 7 presents another multiple judgment method which is considerably more complex than the ''*A*-not *A*.'' Here the subject is presented with a number of *A*'s and an equal number of *B*'s, all unknowns, and his task is to sort them into the two classes. No standard is designated. Figure 7 shows 8 samples, but fewer or more samples may be used. Obtaining the chance probability is not as simple as for the other tests we have discussed. It can be shown that, for 8 samples as shown, the chance probability of a perfect solution is one-thirty-fifth, with higher probabilities for various partially correct sorts *(6)*.

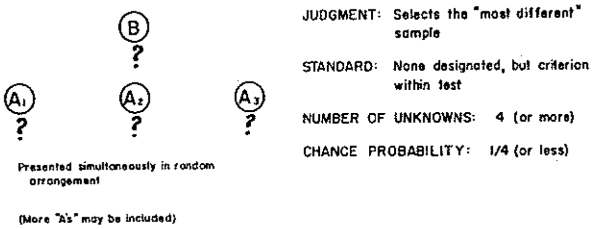
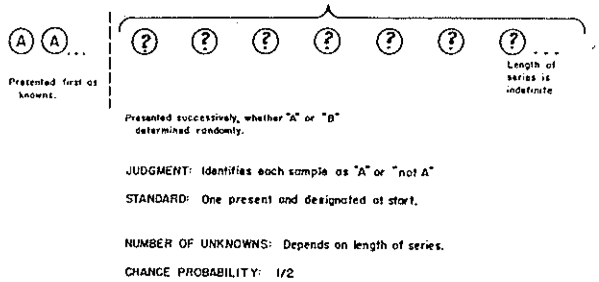

B
?
A_1
?
A_2
?
A_3
?
JUDGMENT: Selects the "most different" sample
STANDARD: None designated, but criterion within test
NUMBER OF UNKNOWNS: 4 (or more)
CHANCE PROBABILITY: 1/4 (or less)
Presented simultaneously in random arrangement
(More "A's" may be included)

Figure 5. Multiple standards test.

Although variations of these test methods are encountered from time to time, they are of only minor importance and do not merit separate discussion because of their close similarity to the seven already described. Understanding and comparison of these methods will be facilitated by noting ways in which they may vary, as follows:

1. *Standards.* The standard may be (a) physically present and designated, as in the duo-trio, dual standard and single stimuli, (b) physically present but not designated, as in the triangle, multiple standards, and multiple pairs, or (c) it may be a designated quality to be remembered by the subject, as in the paired test.


A
A
?
Presented first as knowns.
Presented successively, whether "A" or "B" determined randomly.
Length of series is indefinite

JUDGMENT: Identifies each sample as "A" or "not A"

STANDARD: One present and designated at start.

NUMBER OF UNKNOWNS: Depends on length of series.

CHANCE PROBABILITY: 1/2

Figure 6. *A*-not *A* test (single stimuli).

2. *Statistical efficiency.* As the number of unknowns in the test situation increases the expected chance proportion of correct responses decreases. This variable is often referred to as "statistical efficiency" *(12)*, the test with the lower chance level being considered the more efficient.

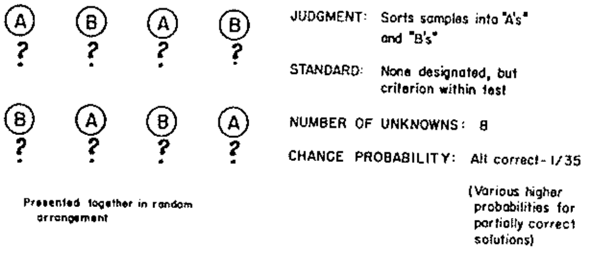
3. *Practical efficiency.* A more practical measure of efficiency is obtained by considering the number of samples involved as well as the chance probability of success on a given try. Some tests give much more information per sample tasted. For example, compare the triangle and single stimuli for 12 samples. With the triangle the chance of all correct responses would be $(\frac{1}{3})^4$ or $\frac{1}{81}$, since there would be four tests. With the single stimuli, allowing for two trials on the known standard, it would be $(\frac{1}{2})^{10}$ or $\frac{1}{1024}$.

4. *Psychological complexity.* The tasks represented in the various tests vary in complexity and difficulty, which, in general, increase with the number of samples which must be considered in arriving at a judgment. The more samples, the greater the possibility of adaptation and interference, and the more demands are placed on the subject's skill and attention.

TEST PROCEDURES

Nearly everything that could be said about necessary physiological and psychological controls in difference testing would be just as true about any sensory testing. However, there are a few points which are especially relevant to difference testing and should be mentioned here.

In tasting a group of samples which constitute a unit, as in the triangle test, there is competition between adaptation and memory. If the time interval between samples is lengthened to permit greater deadaptation it throws a greater burden on flavor memory. Some experimenters attempt to control the interval but more often it is left up to the subject, particularly in triangle testing when samples are presented simultaneously. Experienced panel members seem to be able to work out their own individual methods satisfactorily. Also, it is noted that they tend to allow shorter time intervals than might seem desirable to the novice. It is evident that one should not be arbitrary or dogmatic in regard to these time intervals. Each situation and type of material should be considered on its merits.



JUDGMENT: Sorts samples into "A's" and "B's"

STANDARD: None designated, but criterion within test

NUMBER OF UNKNOWNS: 8

CHANCE PROBABILITY: All correct - 1/35

(Various higher probabilities for partially correct solutions)

Figure 7. Multiple pairs test.

In reviewing the test methods, single test units were described for the most part. However, it is common practice with most of the tests to give a subject more than one test unit at a session. The effect of such multiple testing on sensitivity and discrimination seems to depend on the product. Brandt *(3)* and Mitchell *(14)* have shown that there is no loss through at least 6 successive duo-trios in testing alcoholic beverages and Pfaffmann *(17)* has shown that people can test fruit juices for about 40 minutes without loss of sensitivity. However, for some products, e.g., maple syrup, adaptation takes place rather quickly and no more than one or two test units should be included in a session *(11)*.

It is well to keep in mind that the ostensible standard may not be the only one which is operative and sometimes may not even be the most important. A test subject may develop his own subjective criterion, particularly when considerable testing is done with a constant standard. When this happens in the triangle or duo-trio the third sample is not necessary, and may even be harmful. One could just as well use the simple paired test or the single stimuli. Byer and Abrams *(4)* showed that, when the subject knows the kind of difference to expect, the paired test is superior to the triangle. Raffensperger and Pilgrim *(18)* demonstrated that telling people what sort of a difference to expect improves discrimination in the triangle itself.

Subjects will tend to use all available clues in trying to solve the difference test problem. Even though certain kinds of differences will not help, a person may allow them to affect his judgment. For example, with a solid food there are apt to be many differences among the samples—texture, particle size, shape of portion, etc.—that are purely incidental and have nothing to do with the essential flavor difference which is under test. In such a case the instruction to "pick the different sample" means nothing because all the samples are different. The subject has to know what sort of differences to look for before his response will be meaningful. This is one reason why difference tests are used most often with homogeneous materials.

It is essential to eliminate all incidental differences between the samples that may be correlated with the factor being tested. Otherwise one is likely to misinterpret the results. This caution most often applies when visual differences are associated with possible flavor differences, but are unimportant from the experimenter's standpoint.

PANEL SELECTION AND TRAINING

The assumption is reasonable that it should be possible to select groups of people whose performance on sensory tests will be superior to that of the general population. Various schemes based on intuition, rational judgment, or experimentation have been tried, but with varying degrees of success *(2, 5)*. Further, even though a theoretical system based on experimentation may be available, actual selection may be based on simple practical criteria. One of the major problems is the amount of pretesting work required to effect reliable selection. A further difficulty may be the experimenter's inability to specify accurately the nature of the panel member's task. "Quickie" methods, based upon a few tests, even complex ones, generally have not been very satisfactory. The tedious process of selecting on the basis of sensitivity to the four basic tastes is often recommended, and has even been tried *(13)*, but the method is of doubtful value. Sensitivity is only one factor; successful performance requires such skills as the ability to remember flavors and to compare flavors and flavor strengths in spite of the time lag between samples. Interest and motivation are also very important.

With difference testing, panel selection appears somewhat more feasible than with other types of tests, because an objective criterion of performance is possible. Experience in industry where difference testing is used for flavor quality control on a continuing basis *(10, 15)* has demonstrated the validity of selecting panel members. Here it is feasible because they use panels who test repeatedly, thus providing an adequate sample of each person's behavior. All panel members take essentially the same tests and individual performances are continually evaluated. Those who are right most often stay on the panel, the less skillful are dropped. Potential new members are simply started in the system and are retained only if their relative skill so merits. Thus, in effect, future performance is predicted from past performance.

Although it is commonly assumed that a person needs some training, there is little information available which bears specifically on this point. Industrial experience in flavor quality control work seems to show that training is helpful *(10)*; however, the improvement in over-all panel performance which is usually noted may be due to the selection process rather than to training as such. Another possibility is that a subject's performance will improve because he becomes more familiar with the product being tested, but not because he develops a general skill. Raffensperger and Pilgrim *(18)* found that subjects with no prior experience in difference testing discriminated as well as an experienced panel when appropriate instructions were given. Too little is known to be dogmatic about how much and what kind of training is necessary fully to qualify a panel member. While not discounting the value of having experienced people, one should be aware of the possibility that only a small amount of training may be required in a given instance.

INTERPRETATION

The question is often asked, "How many panel members are needed?" One aspect of this problem is concerned with the reliability of the result. A sufficient number of responses is needed to assure that an important difference will be proved statistically, but this problem is easily solved. One can determine mathematically the number of judgments required to reach any desired level of significance with any given degree of discrimination as shown by the percentage of correct responses. However, the number of judgments can be increased either by using more people or by getting more judgments per person, and this points to the other facet of the problem. The two approaches are not statistically equivalent. As the number of people on the panel is reduced, the generality of the test result is progressively restricted. Perhaps this is of no great concern since the usual difference test panel could hardly represent any definite population anyway. It may be assumed that the population of real interest, about whose behavior we want to make inferences, is that composed of the potential consumers of the product. It is further assumed that the difference test panel is more discriminating than the general consumer—how much more we do not know. Selection and training of panel members and the development of optimum test procedures are designed to assure this; and one can be reasonably certain that a difference not readily apparent to the panel will never be detected by the consumer. This leaves a rather broad range for exercise of the experimenter's judgment; however, in practice, panels are generally small. Panels of less than 5 persons are seldom used, and panels of more than 20 are also rare. Decisions are seldom based on fewer than 16 judgments and it is unusual to obtain more than 30.

Difference test results are limited by the fact that they show only that a particular group of people, at a particular time, and under the conditions of the test, were able to distinguish between two materials, or else failed to prove this within the prescribed test length. Usually we assume that the criterion of difference used by the panel members was related to some physical variable being investigated, but we cannot be sure. Nor are we sure that the difference was the same for all those who responded correctly. Unless there is independent evidence, we do not know whether one sample is better or worse than the other, more or less off-flavor, etc. The exception is the paired test, where a criterion must be specified. Otherwise, all we have is a probability that a discriminable difference exists.

Do the results give degree of difference, as well as the probability of an unknown degree of difference? The answer is "Yes." Intuition tells us that this is true, and so does rational analysis. When we are dealing with a difference in the threshold range, the only appropriate measure of degree of a difference is the frequency with which it is noticed. Thus, an increase in the percentage of correct responses can legitimately be interpreted as indicating greater difference. Note, however, that greater statistical significance resulting from a larger number of responses, with no change in per cent correct, cannot be so interpreted.

APPLICATIONS

One important function is that of controlling the flavor quality of a product against a pre-selected standard. The best examples are in the alcoholic beverage industry, where they have been very effective in achieving product uniformity *(10)*. If product uniformity is a good thing, why don't more companies use the difference testing methods? One answer is that difference tests may impose a more rigid control than many products can afford; also in many instances there is no real need for rigid control. There can be minor variations in a product which would have no practical bearing on consumer acceptance. Here difference-as-such should give way to difference on some continuum such as preference or quality. We want to know not only whether the production lot is different from the standard, but also which is better.

The difference methods are applied extensively in laboratories that do research and development work on foods and beverages *(2)*. A complete review cannot even be attempted here, but they can be applied to such problems as the effect of formulation or processing changes, use of different raw materials, or the effect of packaging materials. The triangle is often used to establish thresholds *(1)*. This is a legitimate use, but the method is cumbersome and inefficient, and the standard psychophysical methods are more suitable.

The triangle test, in particular, has become perhaps a little too famous and has acquired an aura of scientific respectability that may mislead the uniformed. People who are unsophisticated about sensory testing may attempt to use it where it is not appropriate. A frequent error is to attempt to combine a difference test—usually the triangle—with a preference or quality judgment test. If the real problem is something more than simple difference, the information obtained from the difference test does little good. Its inclusion complicates the situation and tends to divert the subject's attention from the dimension of real interest. There is evidence, too, that it will bias the results. Schutz and Bradley have shown that when a preference judgment is asked for, following a triangle test on the same samples, subjects will tend to prefer the paired samples *(19)*.

COMPARATIVE EVALUATION

These difference testing methods are similar in many respects, but there are obvious differences. Some are functionally equivalent; others are not. Although it cannot be precise, an attempt at comparative evaluation will emphasize some important relationships.

Considering the purpose for which they are used, the primary criterion of value for difference tests should be sensitivity, i.e., that test should be considered better which will detect the smaller degree of flavor difference or will demonstrate a given degree of difference with greater certainty, or with less effort. Sensitivity *per se* does not appear to be a very useful criterion. Considerable work has been done to assess relative sensitivity; however, it has not resulted in consistent clear-cut evidence showing the relative merit of the various methods *(17)*. Discrimination, as empirically determined, is too often affected by factors other than the test method itself. Therefore, it will be more useful for our purposes to consider some of the more important subsidiary characteristics, which include efficiency, simplicity and appropriateness to the problem.

Efficiency relates to the amount of work required to achieve a given degree of discrimination. We have already noted that two aspects should be considered. Statistical efficiency is interpreted as the inverse of the chance probability of a correct response, e.g., those tests with a chance probability of $\frac{1}{2}$ would be considered equal and lowest in statistical efficiency whereas the multiple pairs test (Figure 7), with a chance probability of $\frac{1}{35}$, would be considered highest. This aspect has been given more attention than it deserves and has sometimes been erroneously represented as a measure of the sensitivity of a test. In determining practical efficiency one must consider the number of samples tasted in arriving at a judgment as well as the chance probability of a correct judgment. For example, if we equate the total number of samples at 8 or 9 (as appropriate to the test form), then determine the chance probability of all responses being correct, we get the following order of decreasing efficiency: "*A*-not *A*", multiple pairs, triangle, multiple standards and simple paired, duo-trio, and dual standard.

The efficiency of a test, thus defined, cannot be considered independently of simplicity; they interact in determining over-all sensitivity. We are entitled to assume that the simpler test is better, other things being equal. This is a good practical guide even though the superiority of the simpler method may not have been proven in each case. The less the panel member has to draw on complex skills, the less opportunity there is for error, and the less chance for loss of effectiveness due to the selection of inadequate subjects or to the flagging of attention during continued testing. The main factor in complexity is the number of samples, particularly the number of unknowns, that must be considered in arriving at a single judgment. Hence, there is a basic conflict with statistical efficiency. The 7 forms may be ranked as follows in order of decreasing complexity: multiple pairs, multiple standards, triangle, duo-trio, dual standard, single stimuli, and paired. The inversion on the last two is because of the necessity for remembering the standard in the single stimuli.

Appropriateness is a different kind of criterion. It cannot be applied generally, but is specific to the problem and test situation. The necessity of being logically appropriate may restrict the choice of method for certain problems, or a certain method may be found especially suitable for a particular problem. For example, one can not use the paired test unless the dimension and direction of difference can be clearly specified. Again, the multiple standards test is not appropriate except in the unusual situation where the standard is a family of materials rather than a single material. In selecting a test method for regular use, pertinent elements which are anticipated in the test situation should be considered, for example, the probable constancy of the panel, their level of motivation, the amount of testing to be done, whether testing is to be continuous or intermittent, and whether the materials are constant or varied.

Finally, let us try to assign the methods a relative order of merit by consensus, as indicated by their usage. The triangle must be accorded first place because of its wide variety of uses in research, development, and quality control. Next is the duo-trio because of its extensive application in flavor quality control. It could apply wherever the triangle does, although it is not as popular. The paired test deserves third place because of its many research applications, but here the question is always pertinent whether, in any particular test, one is measuring "difference as such" or difference on some particular dimension. When the dimension is a simple one, such as "sweetness," we may call the paired test a difference test; but not when the dimension is a general one, such as "preference" or "quality." The multiple pairs test is ranked fourth because of its frequent appearance in the literature, although usually the emphasis has been on the evaluation of panel members rather than on the testing of materials. It is hardly ever used for anything else. The multiple standards, dual standard, and "*A*-not *A*" tests are tied for last place, since they are seldom used. They are specialized tests and are typical of many other forms which could be developed if needed.

USE OF THE RATING SCALE

The methods described above do not exhaust all possibilities for the measurement of sensory differences. One of the most useful approaches, the rating scale, has been omitted. This basic technique can be applied with many functionally different problems, but is not specifically identified with any one of them. Therefore it can not be considered as belonging to the family of difference tests, although it is often used for difference testing and should be discussed here in order to show its relationship to the other methods.

Such use requires either a verbally designated criterion, as in the paired test, or one which is present and designated, as in the "*A*-not *A*". When using a verbal criterion the restriction again applies that it must be understood by all of the subjects in the same way. Samples are presented singly and the subject indicates the amount or intensity of the designated characteristic on a rating scale. For example, Schutz and Pilgrim *(20)* in measuring differences in sweetness among sugars used a 9-category scale with alternate points anchored as follows: *none, slight, moderate, large, extreme*. This type of judgment is similar to that required in the paired test, where each sample is compared to the other, or in the "*A*-not *A*" method where only the presence or absence of the characteristic is noted, but is more complex than either.

In another application of the rating scale a physical standard is constantly available. The subject considers each unknown in turn, comparing it against the standard and estimating the degree of difference between them. The estimate is recorded on a rating scale like that described above.

The rating scale method is not limited to comparison of only two materials at a time, as are most of the others, but as many as 4 or 5 different samples may be rated at one session. Each subject tests all samples in the group. Successive integral values are assigned to the scale points and the resulting distributions are analyzed by some appropriate technique, such as analysis of variance, to test for the significance of differences between samples.

LITERATURE CITED

1. Berg, H. W., Filipello, F., Hinreiner, E., and Webb, A. D. Evaluation of thresholds and minimum difference concentrations for various constituents of wines: Water solutions of pure substances. *Food Technol.*, 9, 23 (1955).

2. Boggs, Mildred M., and Hanson, Helen. Analysis of foods by sensory difference tests. In *Advances in Food Research*, 2, 220 (1949).

3. Brandt, Donald A., and Hutchison, E. Paul. Retention of taste sensitivity. *Food Technol.*, 10, 419 (1956).

4. Byer, Albert J., and Abrams, Dorothy. A comparison of the triangular and two-sample taste-test methods. *Food Technol.*, 7, 185 (1953).

5. Girardot, Norman F., Peryam, David R., and Shapiro, Ruth. Selection of sensory testing panels. *Food Technol.*, 6, 140 (1952).

6. Gridgeman, N. T. Group size in taste sorting trials. *Food Research*, 21, 534 (1956).

7. Guilford, J. P. *Psychometric Methods.* 2nd ed., 1954. McGraw-Hill, New York, N. Y.

8. Harrison, S., and Elder, L. W. Some applications of statistics to laboratory taste testing. *Food Technol.*, 4, 434 (1950).

9. Helm, Erick, and Trolle, Birger. Selection of a taste panel. *Wallerstein Lab. Commun.*, 9 (28), 181 (1946).

10 Hokenson, Everett P. Upping a food's taste-uniformity. *Food Eng.*, 28, 54 (May 1956).

11 Laue, Elsie A., Ishler, Norman H., and Bullman, Gloria A. Reliability of taste testing and consumer testing methods: Fatigue in taste testing. *Food Technol.*, 8, 387 (1954).

12. Lockhart, Ernest E. Binomial systems and organoleptic analysis. *Food Technol.*, 5, 428 (1951).

13. Mackey, Andrea Overman, and Jones, Patsy. Selection of members of a food tasting panel: Discernment of primary tastes in water solution compared with judging ability for foods. *Food Technol.*, 8, 527 (1954).

14. Mitchell, John W. Duration of sensitivity in trio taste testing. *Food Technol.*, 10, 201 (1956).

15. Peryam, David R. Quality control in the production of blended whiskey. Industrial Quality Control, 7, *3*, 17 (1950).

16. Peryam, David R., and Swartz, Venona W. Measurement of sensory differences. *Food Technol.*, 4, 390 (1950).

17. Pfaffmann, C., Schlosberg, H., and Cornsweet, J. Variables affecting difference tests. In Peryam, D. R., Pilgrim, F. J., and Peterson, M. S. (Eds.), *Food Acceptance Testing Methodology, A Symposium.* Nat. Acad. Sciences-National Research Council, Washington, D. C. 1954, p. 4.

18. Raffensperger, Elsie L., and Pilgrim, Francis J. Knowledge of the stimulus variable as an aid in discrimination tests. *Food Technol.*, 10, 254 (1956).

19. Schutz, H. G., and Bradley, J. Effect of bias on preference in the difference-preference test. In Peryam, D. R., Pilgrim, F. J., and Peterson, M. S. (Eds.), *Food Acceptance Testing Methodology, A Symposium.* Nat. Acad. Sciences-National Research Council, Washington, D. C., 1954, p. 85.

20. Schutz, H. G., and Pilgrim, F. J. The sweetness of various compounds and its measurement. *Food Research*, 21, 206 (1957).